Paper Summarization of the "Interpolating between Optimal Transport and MMD using Sinkhorn Divergences"

Moshood O. Yekini[1], Seedy Jobe[2], Kolawole Tajudeen[3]
Department of Mathematics and Computer Science
African Masters in Machine Intelligence, Accra, Ghana
myekini@aimsammi.org[1], sjobe@aimsammi.org[2], ktajudeen@aimsammi.org@aimsammi.org[3]

1 Introduction and Background of Study

As sample complexity, computational and algorithmic efficiency are prime factors to be considered in modern method designs to be employed in machine learning and imaging sciences for wide ranging applications - shape matching, classification and generative modeling, the notion of comparing of probability measures in addition to the choice of loss function (that ensures good local minima if not guarantee the global minima) in different settings either continuous, discrete or semi-discrete settings has been an area of interest to researchers in the specialization domain.

Quite number of contributions had employed simple norms and the ψ-divergence (being the most commonly used with a classical example of the total variation and Kullback–Leibler divergence) for the comparison of probability densities in point-wise manner. However, due to the mode of operation of this technique, it falls short to capture the geometric property (since it is unstable based on deformations of the distributions' supports) and not in totality obeys the weak convergence of measures.

Definition 1.1 (Weak Convergence) *α_n weakly converges to α, (denoted $\alpha_n \rightarrow \alpha$) $\iff f(x)d\alpha(x)\forall f \in C_b(x)$. Let $\mathcal{D}$ distance between measures, $\mathcal{D}$ metrises weak convergences if and only if*

$$D(\alpha_n, \alpha) \rightarrow 0 \longleftrightarrow \alpha_n \rightharpoonup \alpha$$

For example, on $\mathbb{R}$ with $\alpha = \delta_0$ and $\alpha_n = \delta_{1/n} : D_{kl}(\alpha_n|\alpha)$ there is lack of signal as it approaches the target which causes information loss by the gradient as the $D_{KL}(\alpha_n|\alpha) = +\infty$

To address these drawbacks, other classes of distance metrics are needed to be investigated for which the Maximum Mean Discrepancy (MMD) and the Optimal Transport (OT) distances are found suitable to account for these drawbacks. Also, due to difficulty involve in solving linear problems needed to compute, the entropic regularized OT was considered because of its computational efficiency of approximating the cost function and ease of formulation of a concave dual form of OT with a smooth exponential constraint between sampled measures.

Definition 1.2 (Entropic Regularized Optimal Transport) *Let $\pi \in \mathcal{M}_1^+(\mathcal{X}^2)$ as (π_1, π_2) denotes two marginals of π, for $\epsilon > 0$ the **Entropic Regularized Optimal Transport** distance is defined as*

$$OT_\epsilon(\alpha, \beta) \overset{def.}{=} \underset{\pi_1=\alpha,\pi_2=\beta}{min} \int_{\mathcal{X}^2} C d\pi + \epsilon KL(\pi|\alpha \otimes \beta)$$

$$where\ KL(\pi|\alpha \otimes \beta) \overset{def.}{=} f_{x^2} \log\left(\frac{d\pi}{d\alpha d\beta}\right) d\pi$$

For $C(x,y) = \parallel x - y \parallel^p$ on $\mathcal{X} \subset \mathbb{R}^D$

Definition 1.3 (Reproducing Kernel Hilbert Space) *Let $\mathcal{H}$ a Hilbert space with kernel k, then $\mathcal{H}$ is a **Reproducing Kernel Hilbert Space (RKHS)** if and only if*

- $\forall x \in \mathcal{X}, k(x, \cdot) \in \mathcal{H},$
- $\forall f \in \mathcal{H}, f(x) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}}$

Definition 1.4 (Maximum Mean Discrepancy) *Let $\mathcal{H}$ a RKHS with kernel k, the **Maximum Mean Discrepancy (MMD)** distance between two probability measures α and β is defined by:*

$$\begin{aligned} {MMD_k}^2 &\overset{def.}{=} \left(sup_{f||f||_h \leq 1}|E_\alpha(f(X)) - E_\beta(f(Y)|\right)^2 \\ &= E_{\alpha\otimes\alpha}[k(X,X')] + E_{\beta\otimes\beta}[k(Y,Y')] - 2E_{\alpha\otimes\beta}[k(X,Y)] \end{aligned}$$

Computational Optimal Transport in Machine Learning, African Masters in Machine Intelligence (AMMI 2020), Accra, Ghana.

Considering an Euclidean feature space $\mathcal{X} \subset \mathbb{R}^D$ with Radial Basis Function (RBF) kernel, the kernel loss function (MMD norm) is defined for $\xi = \alpha - \beta$ as

$$L_k(\alpha, \beta) \overset{\text{def.}}{=} \frac{1}{2} \parallel \xi \parallel_k^2 \overset{\text{def.}}{=} \frac{1}{2} \int_{\mathcal{X}^2} k(x, y) d\xi(x) d\xi(y)$$

It has been proved that if the linear space spanned by $k(x, .)$is dense in $\mathcal{C}(\mathcal{X})$ then $\parallel . \parallel_k$ metrizes the convergence in law i.e weak convergence of measure is obeyed.

With this, the MMD norm afford cheaper computation as it scales up to large batches and smaller sample complexity than OT, though the OT assures better gradient of the θ-parameterized loss function for its descent algorithm.

2 Motivation of Study

In order to leverage on the strength of both MMD and OT, there is a need to consider a new cost function of the OT which ensures alternating between optimizing over u with fixed v and optimizing over v with fixed u in the dual formulation of the OT distances untill convergence. This new cost function is called the **Sinkhorn divergence** - a debaised regularized OT distance.

Definition 2.1 (Sinkhorn Divergence) *Let $\alpha \in \mathcal{M}_1^+(\mathcal{X})$ and $\beta \in \mathcal{M}_1^+(\mathcal{Y})$,*

$$S_\epsilon(\alpha, \beta) = OT_\epsilon(\alpha, \beta) - \frac{1}{2} OT_\epsilon(\alpha, \alpha) - \frac{1}{2} OT_\epsilon(\beta, \beta)$$

which satisfies $S_\epsilon(\beta, \beta) = 0$ - address the entropic regularized OT bias and interpolate between OT and MMD. Concisely,

$$OT_0(\alpha, \beta) \xleftarrow{0 \leftarrow \epsilon} S_\epsilon(\alpha, \beta) \xrightarrow{\epsilon \to +\infty} \frac{1}{2} \parallel \alpha - \beta \parallel_{-C}^2$$

3 Main Study

The paper primarily focus on establishing that the motivated Sinkhorn divergence is convex, smooth and positive definite loss function that metrize the convergence in law by developing a fundamental theorem that support these claims.

Theorem 1 (Foundational Theorem) *Let $\mathcal{X}$ be a compact metric space with a Lipschitz cost function $\mathcal{C}(x, y)$ that induces, for $\epsilon > 0$, a positive universal kernel $k_\epsilon(x, y) \overset{def.}{=} \exp(-\mathcal{C}(x, y)/\epsilon)$. Then, S_ϵ defines a symmetric positive definite, smooth loss function that is convex in each of its input variables. It also metrizes the convergence in law: for all probability Radon measures α and $\beta \in \mathcal{M}_1^+(\mathcal{X})$*

$$\begin{aligned} 0 = S_\epsilon(\beta, \beta) &\leq S_\epsilon(\alpha, \beta) \\ \alpha = \beta &\Longleftrightarrow S_\epsilon(\alpha, \beta) = 0 \\ \alpha_n \rightharpoonup \alpha &\Longleftrightarrow S_\epsilon(\alpha_n, \alpha) \to 0 \end{aligned}$$

The authors put forth solid argument of propositions (as provided in appendix) as proof of the theorem paying keen interest to mathematical properties - additive constraint uniqueness, existence of optimal measures, continuity and differentiablity and boundedness which ensures that minimization problems with respect to the OT_ϵ is well-posed and defined.

4 Conclusion

In conclusion, the authors' contributions mathematically affirm the robustness of the Sinkhorn Divergences inheriting the geometric properties from Optimal Transport, breaks the curse of dimensionality and computational efficiency for machine learning and imaging sciences tasks.

References

1. Marco C., Lecture slide on Computational Optimal Transport for Machine Learning, 2019/2020 AMMI, Accra/Rwanda
2. Jean F. et al, Interpolating between Optimal Transport and MMD using Sinkhorn Divergences
3. Aude G., Bridging the gap between optimal transport and MMD with Sinkhorn divergences